Министерство образования и науки РФ

Федеральное государственное бюджетное образовательное учреждение высшего образования

«Омский государственный технический университет»

Кафедра «Автоматизированные системы обработки информации и управления»

Курсовой проект по дисциплине

«Динамические языки программирования»

Выполнил:

Студент гр. ПИН-202 Галинуров А.Е.

(подп., дата)

Проверил:

Доцент каф. АСОИУ

Кабанов А.А.

(подп., дата)

Омск 2023

**Реферат**

ОТЧЕТ 27 с., 22 рис., 3 источника.

КУРСОВОЙ ПРОЕКТ, МАШИННОЕ ОБУЧЕНИЕ

Цель курсовой работы — ознакомиться и приобрести базовые знания в области машинного обучения.

В данной работе выполнены:

1. метод к- ближайших соседей;
2. метод машины опорных векторов;
3. методы линейной и логистической регрессий;
4. метод наивного Байеса;
5. методы решающего дерева и случайного леса;
6. метод CatBoost.

СОДЕРЖАНИЕ

[1. Метод к-ближайших соседей (k-NN) 6](#_Toc154572124)

[1.1 Описание метода 6](#_Toc154572125)

[1.2 Принцип работы 6](#_Toc154572126)

[1.3 Применение на модельных данных 6](#_Toc154572127)

[1.4 Преимущества и ограничения 11](#_Toc154572128)

[2. Метод машины опорных векторов (SVM) 12](#_Toc154572129)

[2.1 Описание метода 12](#_Toc154572130)

[2.2 Разделительная гиперплоскость и принцип работы 12](#_Toc154572131)

[2.3 Применение на модельных данных 12](#_Toc154572132)

[2.4 Преимущества и ограничения 13](#_Toc154572133)

[3. Методы линейной и логистической регрессии 15](#_Toc154572134)

[3.1 Описание методов 15](#_Toc154572135)

[3.2 Принцип работы 16](#_Toc154572136)

[3.3 Сравнительный анализ с другими методами 18](#_Toc154572137)

[4. Метод наивного Байеса 19](#_Toc154572138)

[4.1 Описание метода 19](#_Toc154572139)

[4.2 Применение на модельных данных 19](#_Toc154572140)

[4.3 Преимущества и ограничения 20](#_Toc154572141)

[5.1 Описание метода 21](#_Toc154572142)

[5.2 Применение на модельных данных 21](#_Toc154572143)

[5.3 Сравнение с другими методами классификации 23](#_Toc154572144)

[6. Метод CatBoost 24](#_Toc154572145)

[6.1 Описание метода 24](#_Toc154572146)

[6.2 Применение на модельных данных 24](#_Toc154572147)

[Заключение 26](#_Toc154572148)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 27](#_Toc154572149)

**Введение**

Машинное обучение — одна из наиболее динамично развивающихся областей, которая позволяет компьютерам обучаться на основе данных и делать прогнозы или принимать решения без явного программирования.

Цель настоящей курсовой работы заключается в ознакомлении с основами машинного обучения и рассмотрении различных методов классификации. В ходе исследования были проведены эксперименты с различными алгоритмами, такими как метод к-ближайших соседей, машина опорных векторов, линейная и логистическая регрессии, наивный Байес, а также решающее дерево и случайный лес.

В рамках данной работы представлен обзор и сравнение указанных методов классификации на простых модельных данных. Каждый из методов исследован с целью понимания их принципов работы, преимуществ и ограничений.

# Метод к-ближайших соседей (k-NN)

# Описание метода

Метод к-ближайших соседей (k-NN) относится к одному из простейших алгоритмов классификации в машинном обучении. Он основывается на принципе близости объектов: если у объекта есть соседи известного класса, то скорее всего, этот объект также принадлежит к этому классу.

# Принцип работы

При классификации нового объекта метод k-NN ищет k ближайших к нему объектов в обучающем наборе данных. Затем присваивает новому объекту тот класс, который наиболее часто встречается среди его соседей.

# Применение на модельных данных

На рисунке 1 представлен датасет, на основе которого будут применяться все представленные методы:

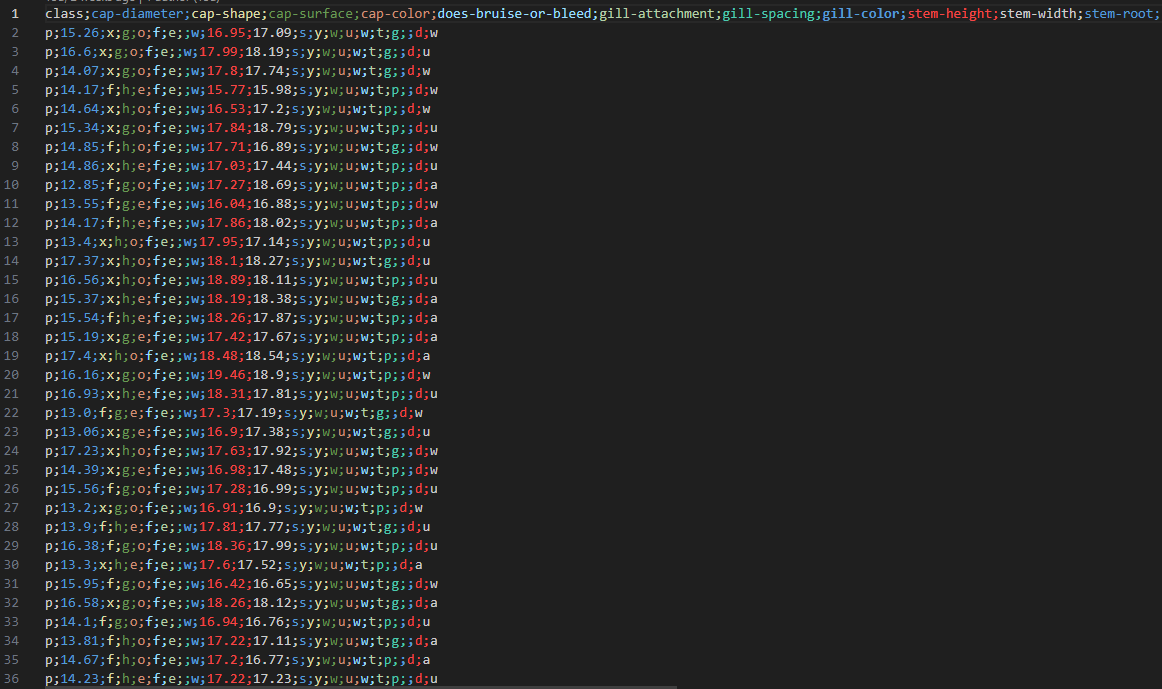


Рисунок 1 – Датасет

Набор данных смоделированных грибов для бинарной классификации на съедобные и ядовитые. В каждой строке есть информация о грибе, представленная различными атрибутами:

1. **CAP-DIAMETER**: диаметр колпачка (m): число плавучести в см.
2. **CAP-SHAPE**: форма крышки (n): колокольчатая=b, коническая=c, выпуклая=x, плоская=f, утопленная=s, сферическая=p, другие=o
3. **CAP-SURFACE**: поверхность шляпки (n): волокнистая=i, бороздки=g, чешуйчатая=y, гладкая=s, блестящая=h, кожистая=l, шелковистая=k, липкая=t, морщинистый=w, мясистый=e
4. **CAP-COLOR**: цвет шапки (n): коричневый=n, буфф=b, серый=g, зеленый=r, розовый=p, фиолетовый=u, красный=e, белый=w, желтый=y, синий=l, оранжевый=o, черный=k
5. **DOES-BRUISE-BLEED**: кровоточит ли синяк (n): синяки или кровотечение=t, нет=f
6. **GILL-ATTACHMENT**: жаберный аппарат (n): аднатный=a, аднексированный=x, декуррентный=d, свободный=e, синуат=с, поры=п, нет=ф, неизвестно=?
7. **GILL-SPACING** расстояние между жабрами (n): близко=c, далеко=d, нет=f
8. **GILL-COLOR**: цвет жабр (n): см. цвет крышечки + нет=f
9. **STEM-HEIGHT**: высота стебля (м): плавающее число в см.
10. **STEM-WIDTH**: ширина стебля (м): плавающее число в мм
11. **STEM**-**ROOT**: стеблекорень (n): луковичный=b, вздутый=s, булавовидный=c, чашевидный=u, равный=e, ризоморфы=z, корневище=r
12. **STEM-SURFACE**: поверхность стебля (n): см. поверхность шапки + нет=f
13. **STEM-COLOR**: цвет стебля (n): см. цвет шапки + none=f
14. **VEIL-TYPE**: тип вуали (n): частичная=p, универсальная=u
15. **VEIL-COLOR**: вуаль-цвет (n): см. колпачок-цвет + none=f
16. **HAS-RING**: имеет ли кольцо? Кольцо=t, нет=f
17. **RING-TYPE**: тип кольца (n): паутинистое=c, эмансипирующее=e, вспыхивающее=r, рифленое=g, большое=l, подвесное=p, оболочка=s, зональное=z, чешуйчатое=y, подвижное=m, нет=f, неизвестное=?
18. **SPORT-PRINT-COLOR**: цвет отпечатка споры (n): см. цвет колпачка
19. **HABITAT**: среда обитания (n): травы=g, листья=l, луга=m, тропинки=p, пустоши=h, городские=u, отходы=w, леса=d
20. **SEASON**: сезон (n): весна=s, лето=u, осень=a, зима=w
21. **CLASS:** целевая переменная, где p указывается на то, что гриб является ядовитым, а e ­ съедобным.

На рисунках 2–4 представлена подготовка данных

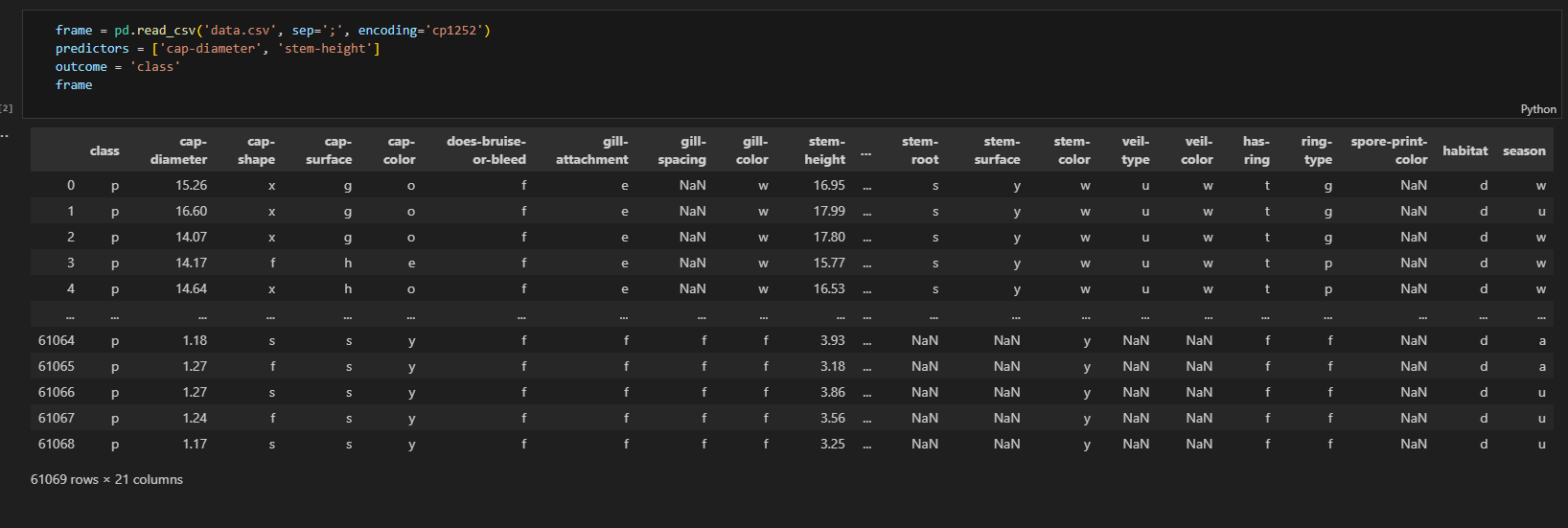


Рисунок 2 – Подготовка данных



Рисунок 3 – Подготовка данных

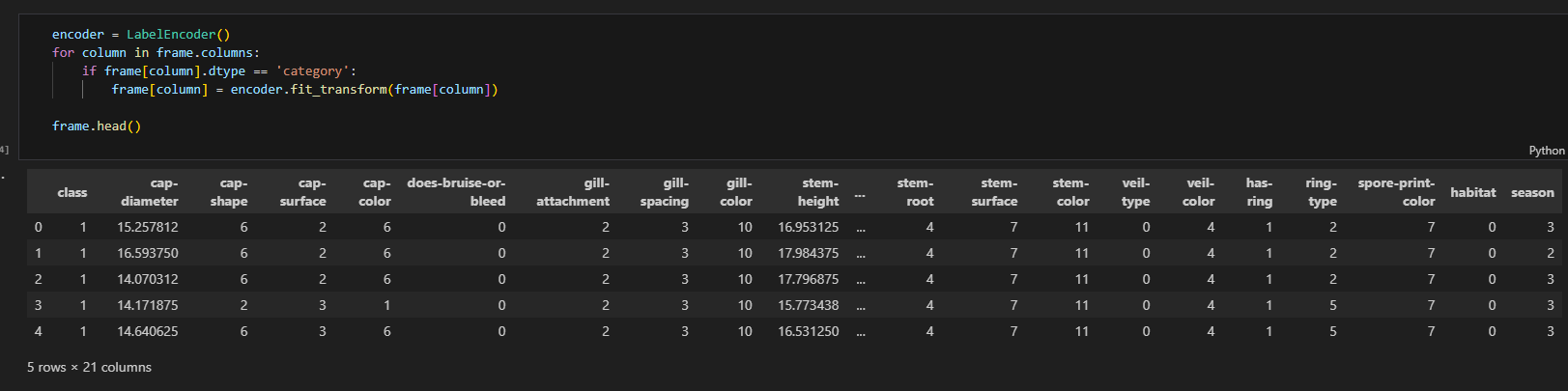


Рисунок 4 – Подготовка данных

На рисунке 5 представлены разделение данных и нормализация

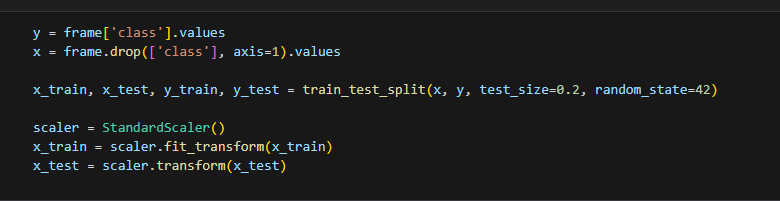


Рисунок 5 – Разделение данных и нормализация

На рисунке 6 представлены подбор гиперпараметров и наилучшего k

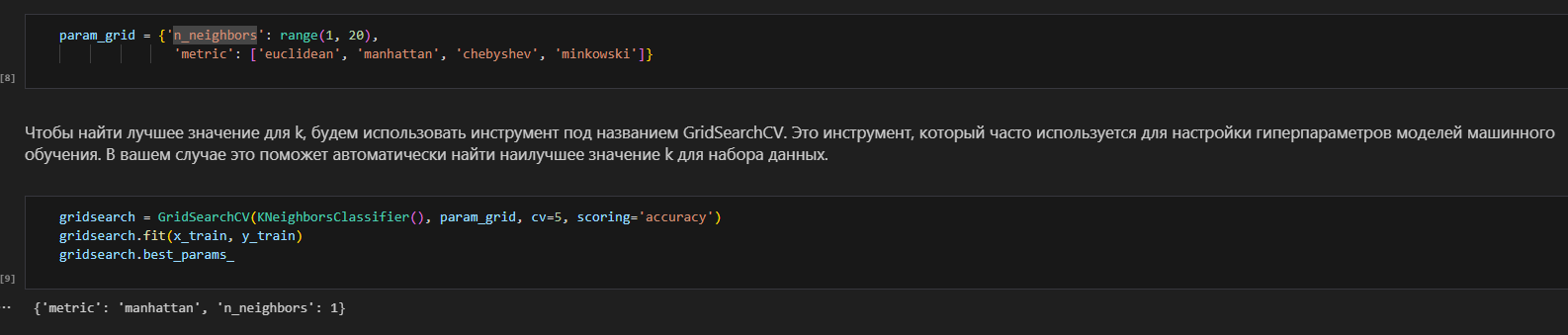


Рисунок 6 – Подбор гиперпараметра и наилучшего k

На рисунке 7 представлены проверка на тестовых данных и вывод результатов

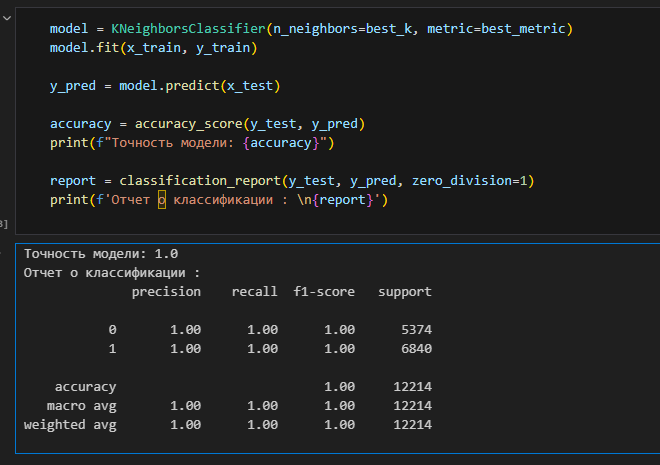


Рисунок 7 – Проверка на тестовых данных и вывод результатов

# Преимущества и ограничения

Преимущества метода k-NN:

* Прост в реализации и понимании.
* Не требует обучения модели, пока поступают новые данные.
* Хорошо подходит для начальной оценки данных и быстрых прототипов.

Однако у метода k-NN есть и ограничения:

* Чувствителен к выбросам в данных.
* Неэффективен на больших наборах данных из-за вычислительной сложности.
* Не учитывает значимость признаков, все признаки равнозначны.

Метод k-NN полезен для первичного анализа данных, но в реальных приложениях может потребоваться более сложная модель для учета различных аспектов и повышения точности предсказаний.

# Метод машины опорных векторов (SVM)

# Описание метода

Метод машины опорных векторов (SVM) является мощным алгоритмом машинного обучения, используемым для задач классификации и регрессии. Основная идея заключается в поиске оптимальной разделительной гиперплоскости, которая максимально разделяет классы в данных.

# Разделительная гиперплоскость и принцип работы

SVM строит гиперплоскость в n-мерном пространстве, где n - количество признаков. Эта гиперплоскость разделяет пространство на две части и максимизирует расстояние (зазор) между объектами разных классов, называемое отступом. Оптимальная гиперплоскость выбирается так, чтобы этот отступ был максимальным.

# Применение на модельных данных

На рисунке 8 представлены гиперпараметры и подбор гиперпараметров с помощью перекрестной проверки

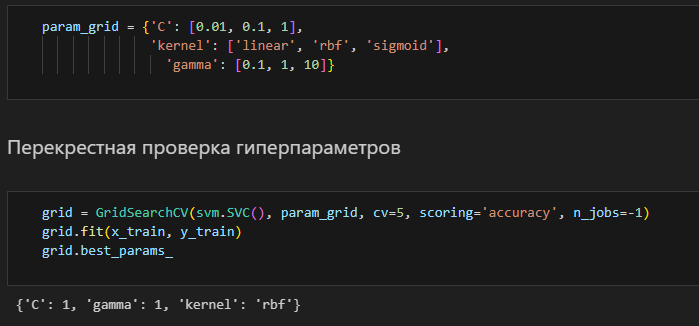


Рисунок 8 – Подбор гиперпараметров с помощью перекрестной проверки

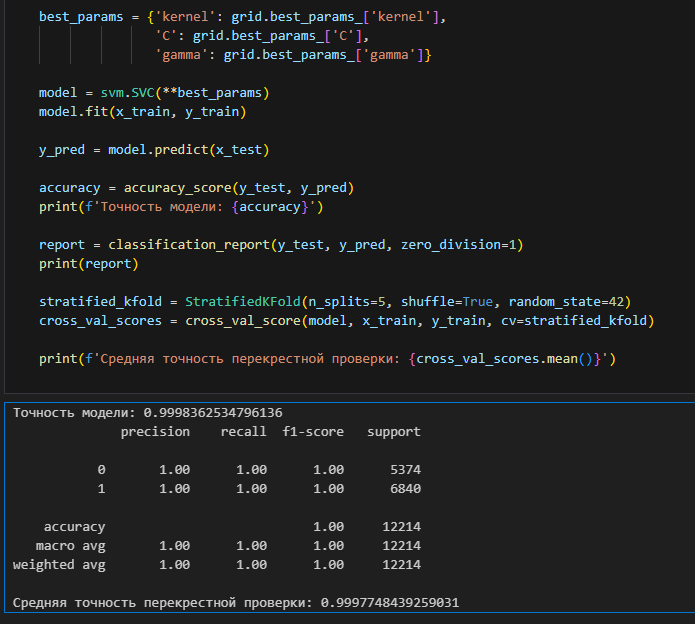
Вывод лучших гиперпараметров и оценка производительности модели на тестовом наборе

Рисунок 9 ­ Оценка производительности модели на тестовом наборе

# Преимущества и ограничения

Преимущества метода SVM:

* Эффективен в пространствах с большим количеством признаков.
* Хорошо работает в условиях разделяемых данных с помощью нелинейных ядер.
* Стабилен при обучении на небольшом наборе данных.

Однако у метода SVM есть и ограничения:

* Требует тщательного выбора гиперпараметров и ядра.
* Неэффективен при работе с большими наборами данных из-за вычислительной сложности.
* Может быть чувствителен к выбросам в данных.

SVM - мощный алгоритм, который может быть эффективен при правильной настройке, но требует опыта для правильного применения и настройки гиперпараметров для достижения оптимальных результатов.

# Методы линейной и логистической регрессии

# Описание методов

Линейная регрессия – это метод, используемый для прогнозирования значений непрерывной зависимой переменной на основе линейной комбинации независимых переменных. Основная идея заключается в поиске линейной зависимости между предикторами и целевой переменной.

Применение линейной регрессии может быть полезным при прогнозировании численных значений, например, прогнозировании цены на недвижимость на основе её характеристик, таких как площадь, количество комнат и т.д.

Логистическая регрессия используется для задач классификации, прогнозируя вероятность принадлежности объекта к определенному классу. В отличие от линейной регрессии, логистическая регрессия применяется для бинарной или многоклассовой классификации.

# Принцип работы

На рисунках 10 – 11 представлена работа с линейной регрессией.

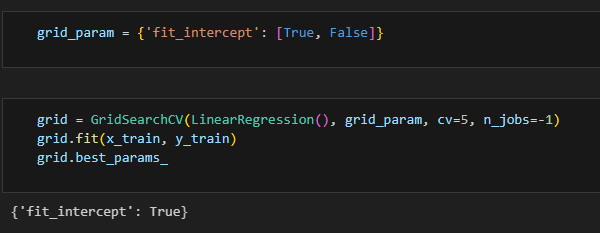


Рисунок 10 – Перекрестная проверка гиперпараметров линейной регрессии

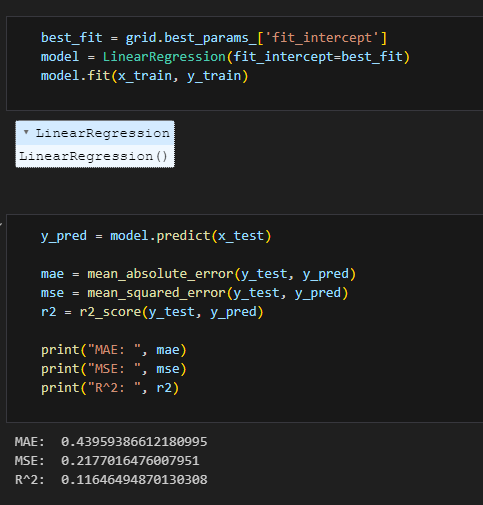


Рисунок 11 – Оценка точности модели на тестовых данных

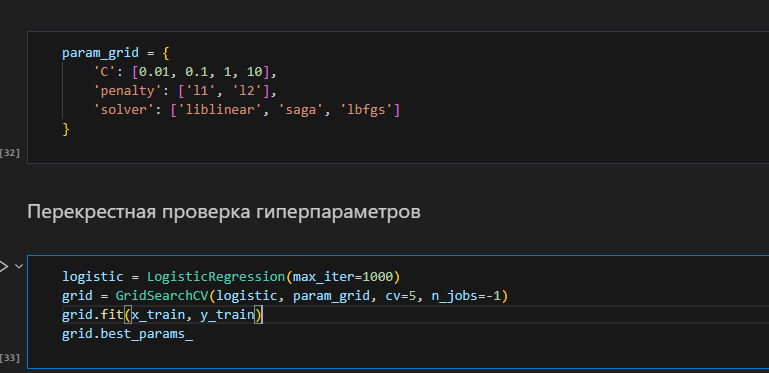
На рисунках 12 – 13 представлена работа с логистической регрессией.

Рисунок 12 ­ Перекрестная проверка гипермараметров логистической регрессии

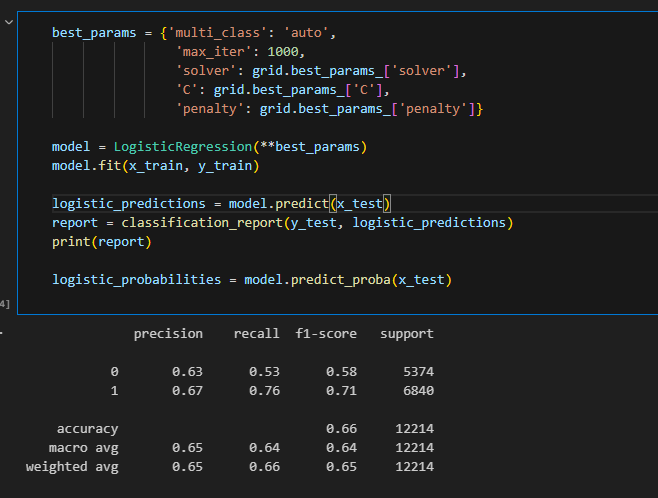


Рисунок 13 – Оценка точности модели на тестовых данных

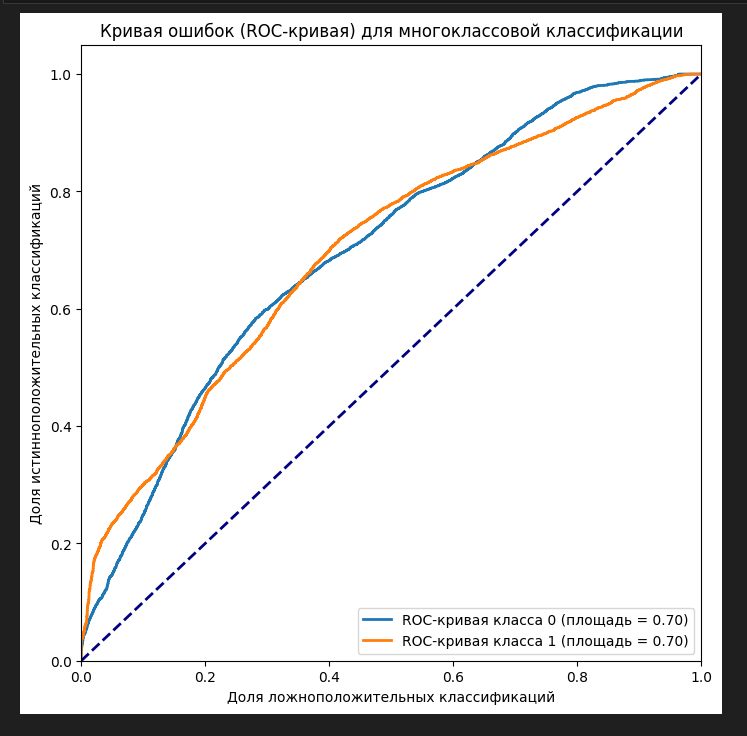


Рисунок 14 – Кривая ошибок для многоклассовой классификации

# Сравнительный анализ с другими методами

Линейная и логистическая регрессии имеют свои сильные и слабые стороны по сравнению с другими методами. Линейная регрессия хорошо работает при предсказании непрерывных значений, тогда как логистическая регрессия применяется для задач классификации. Однако оба метода чувствительны к выбросам в данных и могут быть недостаточно гибкими для улавливания сложных нелинейных зависимостей. Поэтому в случае сложных данных и нелинейных связей эти методы могут оказаться менее эффективными по сравнению с другими алгоритмами, такими как деревья решений или нейронные сети.

# Метод наивного Байеса

# Описание метода

Метод наивного Байеса основан на теореме Байеса и предполагает независимость между признаками. Он использует вероятностный подход к классификации и основывается на простой предпосылке, что признаки объектов независимы между собой при условии принадлежности к определенному классу.

# Применение на модельных данных

На рисунке 15 представлен поиск лучших гиперпараметров и модели наивного Байеса

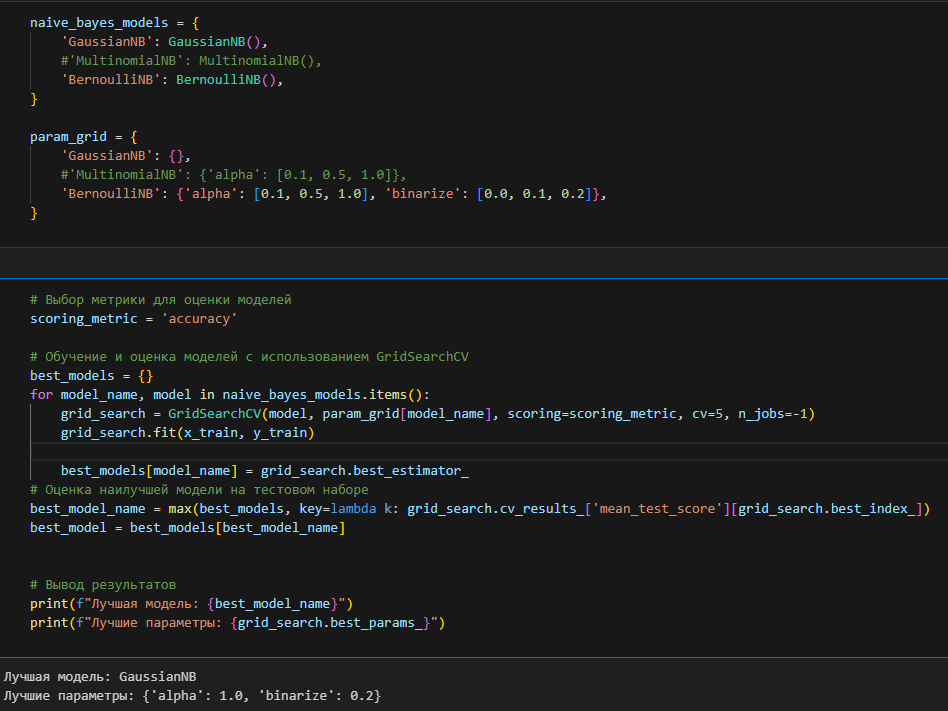


Рисунок 15 – Перекрестная проверка гиперпараметров

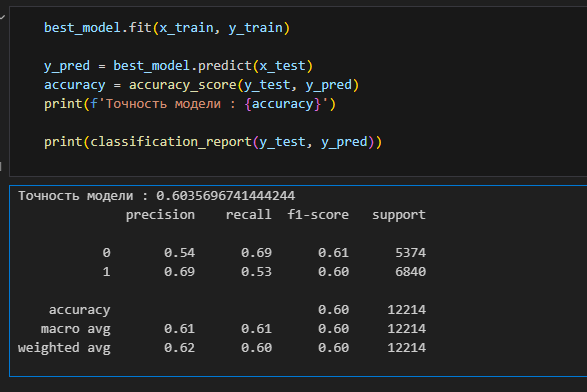


Рисунок 16 – Оценка точности модели

# Преимущества и ограничения

Преимущества метода наивного Байеса:

* Эффективен и быстр в обучении, особенно при работе с большими наборами данных.
* Хорошо работает в условиях небольшой обучающей выборки.
* Прост в реализации и понимании.

Однако у метода наивного Байеса есть и ограничения:

* Предполагает независимость признаков, что может быть не всегда верным в реальных данных.
* Может давать неоптимальные результаты в случае, когда зависимости между признаками сильны.
  1. **Методы решающего дерева и случайного леса**

# Описание метода

Решающее дерево представляет собой структуру, состоящую из узлов и листьев, которая используется для принятия решений на основе вопросов о значениях признаков. Оно строит дерево, разбивая данные на подмножества на основе определенных признаков.

Принцип работы решающего дерева заключается в выборе наилучшего признака для разделения данных на каждом узле дерева. Этот процесс повторяется, пока не будет достигнуто условие остановки, такое как максимальная глубина дерева или минимальное количество объектов в листе.

# Применение на модельных данных

-

На рисунке 17 представлен метод решающего дерева

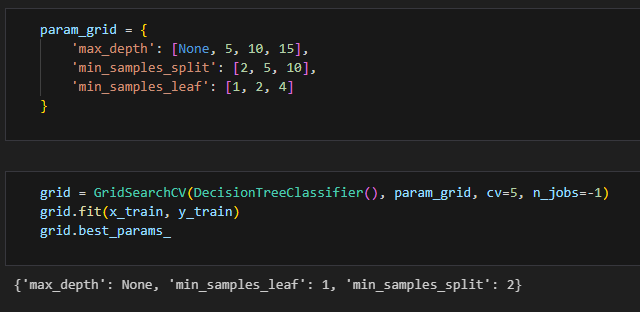


Рисунок 17 – Метод решающего дерева

На рисунке 18 представлены обучение решающего дерева и вывод результата

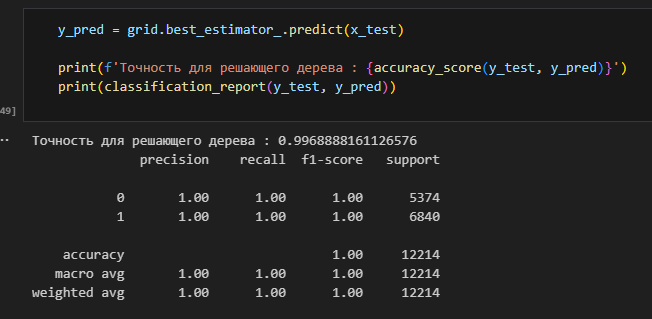


Рисунок 18 – Обучение и вывод результата

На рисунке 19 представлен метод случайного дерева

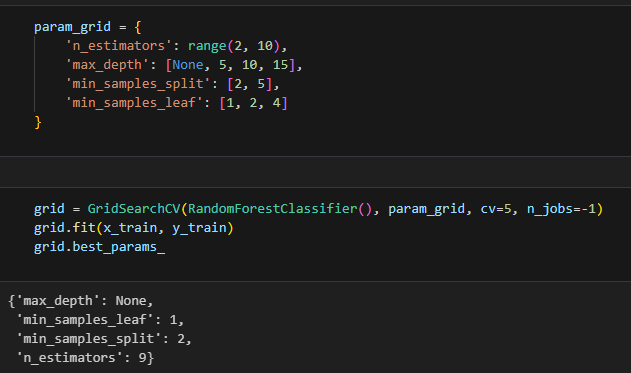


Рисунок 19 ­ Метод случайного дерева

На рисунке 20 представлены обучение случайного дерева и вывод результата.

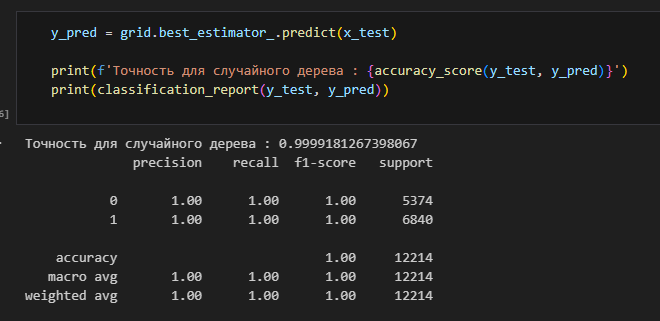


Рисунок 20 – Обучение и вывод результата

# Сравнение с другими методами классификации

Решающее дерево и случайный лес являются эффективными методами классификации, которые обладают рядом преимуществ:

* Способны обрабатывать как категориальные, так и числовые данные.
* Позволяют визуализировать принятие решений.
* Могут обрабатывать большие объемы данных.

Однако у этих методов также есть некоторые ограничения:

* Склонны к переобучению на обучающих данных.
* Могут не улавливать сложные нелинейные зависимости.

По сравнению с линейными моделями, решающие деревья и случайный лес могут быть более гибкими в обработке сложных данных, но требуют осторожной настройки параметров для предотвращения переобучения. Их эффективность часто зависит от природы данных и особенностей задачи классификации.

# Метод CatBoost

# Описание метода

CatBoost (Categorical Boosting) – это высокоэффективная библиотека градиентного бустинга, специально разработанная для работы с категориальными признаками. Она представляет собой мощный алгоритм машинного обучения, который широко применяется в задачах классификации, регрессии и ранжирования.

# Применение на модельных данных

На рисунке 21 представлена реализация метода CatBoost

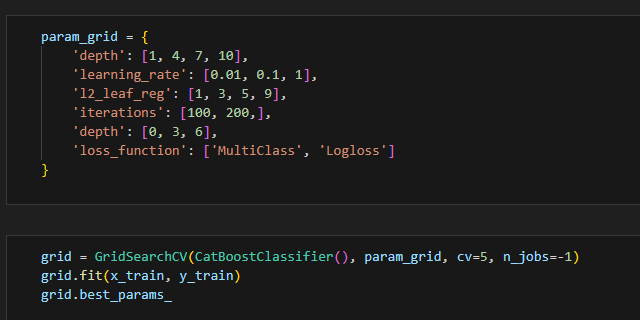


Рисунок 21 – Метод CatBoost

На рисунке 22 преставлен вывод результата работы метода CatBoost

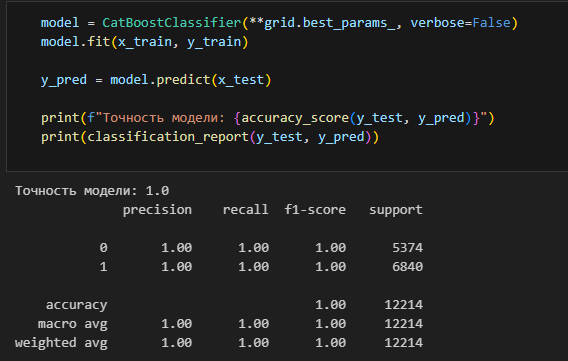


Рисунок 22 – Вывод результата

CatBoost обладает мощными возможностями для работы с категориальными данными и является одним из популярных алгоритмов для задач машинного обучения, особенно в случаях, когда требуется обработка сложных и разнородных данных без предварительной подготовки.

# Заключение

В рамках данного исследования были рассмотрены и проанализированы различные методы машинного обучения, каждый из которых представляет собой мощный инструмент для решения задач классификации и регрессии.

Выбор конкретного метода зависит от природы данных, размера выборки, задачи и требований к точности предсказаний. Каждый метод имеет свои преимущества и ограничения, и правильный выбор может существенно повлиять на результаты анализа. Для повышения эффективности и точности модели часто требуется комбинирование нескольких методов или тщательная настройка параметров.

В целом, эта работа служит важным введением в мир машинного обучения и предоставляет базовые знания о различных методах, что поможет в выборе и применении наиболее подходящего метода для конкретной задачи.

# СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Основы машинного обучения, лекция 2 — метод k ближайших соседей: [https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCR](https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=2) [xBzxsq9lkJkzMgzWiyg&index=2](https://www.youtube.com/watch?v=X081VuXB1og&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=2)
2. Основы машинного обучения, лекция 4 — линейная регрессия: [https://www.youtube.com/watch?v=8RAXDT\_5\_js&list=PLEwK9wdS5g0oCR](https://www.youtube.com/watch?v=8RAXDT_5_js&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=24) [xBzxsq9lkJkzMgzWiyg&index=24](https://www.youtube.com/watch?v=8RAXDT_5_js&list=PLEwK9wdS5g0oCRxBzxsq9lkJkzMgzWiyg&index=24)
3. Андрей Бурков. Машинное обучение без лишних слов